FlyCyc: updating the metabolic network for *Drosophila melanogaster*

Steven J Marygold[1*], Phani V Garapati[1], Gil dos Santos[2] and Peter D Karp[3]

1. FlyBase, Department of Physiology, Development and Neuroscience, University of Cambridge, Cambridge, UK
2. FlyBase, The Biological Laboratories, Harvard University, Cambridge, MA, USA
3. Bioinformatics Research Group, SRI International, Menlo Park, CA, USA

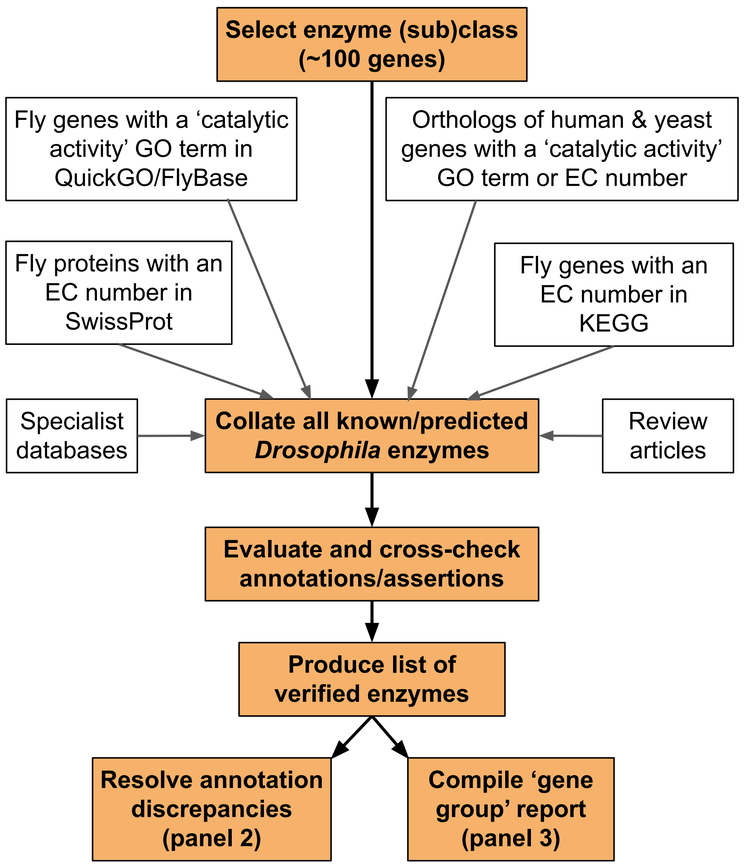
* email: **sjm41@cam.ac.uk**

BioCyc is a collection of metabolic networks for over 20,000 species. The BioCyc 'Pathway Tools' software can generate a metabolic network by matching reactions/pathways in its reference database (MetaCyc) with the set of enzymes encoded by given genome. The quality of the metabolic reconstruction therefore depends on the accuracy and completeness of enzymatic annotation of that genome. A BioCyc model for *Drosophila melanogaster* (FlyCyc) exists but is based on incomplete data from a FlyBase release 15 years ago and therefore does not include more recent improvements to functional annotations.

We have conducted a systematic review of *Drosophila* enzymes, improving the coverage and accuracy of their functional annotations (Gene Ontology (GO) and Enzyme Commission (EC)) in FlyBase. Overall, we verified ~3,750 Drosophila enzymes and made ~4,000 changes to manual annotations. We have also improved access to enzymatic data within FlyBase by displaying EC information and chemical reaction graphics (from the RHEA database) in relevant gene reports, and by creating accessible 'gene group' pages representing each enzyme class/subclass.

The revisions to enzyme annotations have allowed us to compute a new FlyCyc that also incorporates the latest genomic and gene nomenclature data. Compared to the previous version, the updated FlyCyc includes >50 additional metabolic pathways and identifies >600 additional enzyme-encoding genes. However, a number of ambiguous enzyme mappings and 'pathway holes' remain - as far as possible, these are being resolved by correcting GO/EC annotations.. Once finalized, the new FlyCyc will be made available on the BioCyc website and via FlyBase, thereby providing researchers with much-improved *Drosophila* metabolic pathway diagrams and enhanced capabilities to analyse metabolomic datasets.

1. Review *Drosophila* enzymes



2. Resolve annotation discrepancies

Enzyme class	Number of annotated genes*		
	before review	after review	removed/added
Oxidoreductases	617	621	88 / 92
Transferases	1,317	1,301	222 / 206
Hydrolases	1,781	1,567	440 / 226
Lyases	119	133	13 / 27
Isomerases	96	104	9 / 17
Ligases	111	146	16 / 51
Translocases	133	142	44 / 53
TOTAL	4,174	4,014	832 / 672

* Number of genes annotated to corresponding GO terms in FB2017_05 cf FB2023_02

Causes of annotation discrepancies:

- Erroneous/missing computational GO annotation
- Erroneous/missing EC xref in the GO
- Erroneous/missing relationship in the GO
- Erroneous/missing manual GO annotation
- Uncurated literature
- Lack of equivalence between GO and EC
- Database asynchrony
- No GO term
- Incorrect EC annotations submitted to INSDC
- Erroneous/missing EC/keyword in Swiss-Prot

3. Compile verified enzymes as 'Gene Groups'

FlyBase 'Gene Groups' are manually-curated collections of functionally-related *D. melanogaster* genes. They are arranged into hierarchies, cross-referenced with applicable GO (and EC) terms, and provide links to relevant literature, FlyBase tools and equivalent groups of human genes at the HGNC. Our organization of enzyme gene groups follows that of the EC/GO.

General Information

Name	L-MALATE DEHYDROGENASES	Species	*D. melanogaster*
Symbol	LMDH	FlyBase ID	FBgg0001710
Date last reviewed	2021-03-18	Number of members	4

Description

Description	L-malate dehydrogenases are NAD/NADH-dependent oxidoreductases that catalyze interconversion of the substrates malate and oxaloacetate. This reaction plays key role in the malate/aspartate shuttle across the mitochondrial membrane, and in the tricarboxylic acid cycle within the mitochondrial matrix. (Adapted from PMID:12537350).
Notes on Group	
Source Material	The L-MALATE DEHYDROGENASES Gene Group has been compiled using the following publication(s): *Voelker et al., 1979* .

Key Gene Ontology (GO) terms

Molecular Function	L-malate dehydrogenase activity
Biological Process	tricarboxylic acid cycle
Cellular Component	mitochondrion

Enzymatic activity

Enzyme name (EC)	malate dehydrogenase (1.1.1.37)

Related Gene Groups

Parent group(s)	MALATE DEHYDROGENASES

Members (4)

For all members: View Orthologs | Export to HitList | Export to Batch Download

GO ribbon stack

Gene Symbol	Gene Name	Also Known As	Source Material for Membership
CG10748			(*FlyBase, 2017-*)
CG10749			(*FlyBase, 2017-*)
Mdh1	Malate dehydrogenase 1	Mdh, Mdh-1, cMdh, Malate dehydrogenase	(*FlyBase, 2017-*, *Voelker et al., 1979*)
Mdh2	Malate dehydrogenase 2	Malate dehydrogenase, psg7, l(3)psg7	(*FlyBase, 2017-*, *Voelker et al., 1979*)

External Data

Equivalent Group(s)	
Other resource(s)	

Synonyms and Secondary IDs

References (3)

4. Improve enzyme data on Gene Reports

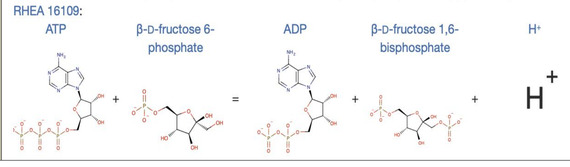
The EC name and number now appear in the General Information section of relevant Gene Reports. The EC reaction description and a reaction graphic from RHEA are shown in the Function section. EC/RHEA annotations are computed from our GO annotations.

General Information

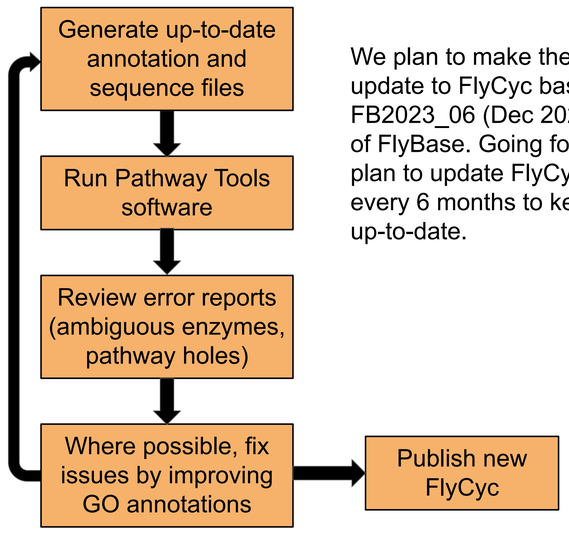
Symbol	Dmel\Pfk	Species	*D. melanogaster*
Name	Phosphofructokinase	Annotation Symbol	CG4001
Feature Type	protein_coding_gene	FlyBase ID	FBgn0003071
Gene Model Status	Current	Stock Availability	12 publicly available
Enzyme Name (EC)	6-phosphofructokinase (2.7.1.11)		

Function

Catalytic Activity (EC/Rhea)	**6-phosphofructokinase activity** ATP + beta-D-fructose 6-phosphate = ADP + beta-D-fructose 1,6- bisphosphate + H(+) (2.7.1.11) RHEA 16109:



5. Update FlyCyc



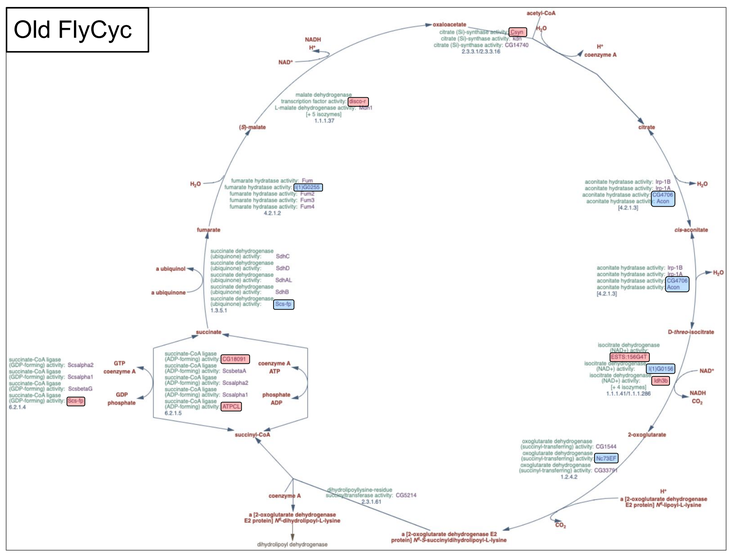
We plan to make the first public update to FlyCyc based on the FB2023_06 (Dec 2023) release of FlyBase. Going forwards, we plan to update FlyCyc at least every 6 months to keep the data up-to-date.

6. FlyCyc improvements

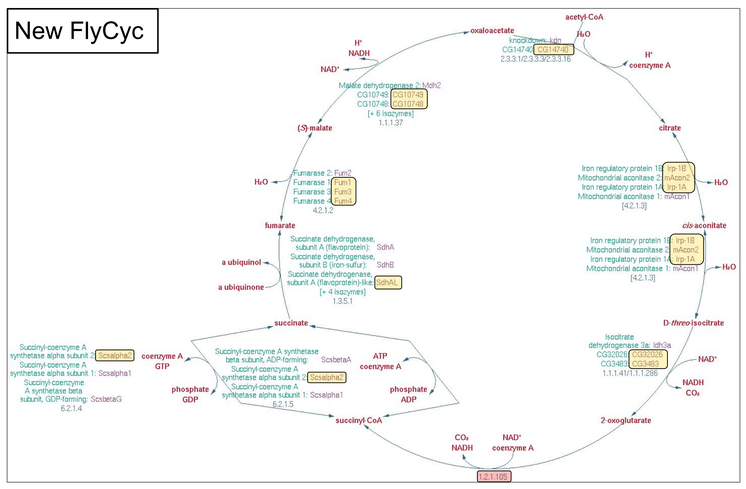
Summary statistics illustrating the main differences between the old and new versions of FlyCyc are shown on the right. (The total number of enzymes has decreased mainly because many genes were wrongly annotated with catalytic GO terms in the past.) Differences in the TCA (Krebs, citric acid) cycle are shown below as a specific example. The old pathway diagram (left) contains several errors (red highlight) and displays old gene nomenclature (blue highlight). The newly computed diagram (right) is much better, though it still doesn't differentiate between canonical members and testis-specific factors (yellow highlight).

	Old FlyCyc	New FlyCyc
Genome release	5.10	6.54
FlyBase release	FB2008_07	FB2023_05
Total #genes	15,097	17,856
Total #enzymes	3,504	2,342
Total #pathways	230	288
Total #GO terms	170	91,329

Old FlyCyc



New FlyCyc



7. Future plans

We will use the newly computed FlyCyc pathways, together with empirical data from published papers and computed pathway information available at Reactome and KEGG, to create a set of high-quality, manually-curated metabolic pathways for *D. melanogaster*. We will use the Gene Ontology Causal Activity Model (GO-CAM) framework that is being adopted across the model organism databases. GO-CAMs are based directly on GO annotations created and maintained by FlyBase curators and can accommodate tissue- and context-specific pathways, such as testis-specific pathways. The GO-CAM models will be published within new Metabolic Pathway Reports at FlyBase and on the Alliance of Genome Resources website.

Acknowledgments: We thank Pascale Gaudet, Harold Drabkin, Marc Feuermann, the InterPro curators and other members of the GO consortium for addressing the hundreds of GO tickets and disputes raised during this work. Thanks also to Helen Attrill for advice on GO annotation, Kristian Axelson and Ron Caspi for help with EC queries, and David Hill and Peter D'Eustachio for Reactome2GO/GO-CAM discussions. This work is funded by the National Human Genome Research Institute and the National Institute of Diabetes and Digestive and Kidney Diseases at the NIH (#U41HG000739 and 1R01DK136945-01).